Homework #3
CS699 Fall 2017

Due Friday Dec 1 by 5:00pm

General Instructions The following assignment is meant to be challenging. Feel free to discuss with fellow students, though please write up your solutions independently and acknowledge everyone you discussed the homework with on your writeup. We also expect that **you will not attempt to consult outside sources, on the Internet or otherwise,** for solutions to any of these homework problems. Finally, unless otherwise stated please provide a formal mathematical proof for all your claims.

Note To submit the homework, you may either email it to Li, hand it to him in person, or drop it off in the box which we will make available in SAL 246 on Friday. If you need more time, consider using one of your late days.

Problem 1. (20 points)
In this problem, we will relate strictly proper scoring rules to distance functions known as *Bregman Divergences*. Given a convex set Ω and a strictly convex and differentiable function $F : \Omega \to \mathbb{R}$, the Bregman distance from p to q is given by $D_F(p, q) = F(p) - (F(q) + \bigtriangledown F(q) \cdot (p - q))$. In other words, it is the vertical distance, taken at p, between the function F and its first-order Taylor approximation around q. It is easy to verify that $D_F(p, q) \geq 0$, with equality holding if and only if $p = q$, so we can think of D_F as a distance function on Ω. However, D_F is not necessarily a metric, since it fails to satisfy symmetry and the triangle inequality in general. We say a distance function $D : \Omega \times \Omega \to \mathbb{R}_+$ is a Bregman divergence if there exists a strictly convex and differentiable function F such that $D = D_F$.

a. (4 points). Show that the KL-divergence and the squared Euclidean distance are both Bregman divergences.

b. (5 points). Let $D : \Omega \times \Omega \to \mathbb{R}_+$ be a Bregman divergence on a convex set $\Omega \subseteq \mathbb{R}^n$, and let $v \in \Omega$ be a random variable with expectation $\overline{v} \in \Omega$. Show that $\overline{v}$ minimizes expected distance from v, formally $\overline{v} \in \operatorname{argmin}_{u \in \Omega} \mathbf{E}[D(v, u)]$.

We consider scoring rules of the form $S : \Delta_n \times [n] \to \mathbb{R}$. As in class, we extend the domain of a scoring rule S to $\Delta_n \times \Delta_n$ by defining $S(p, q) = \mathbf{E}_{i \sim q} S(p, i)$. Define the *loss function* L_S of a scoring rule S by $L_S(p, q) = S(p, p) - S(q, p)$; recall from class that $L_S(p, q)$ can be thought of as the "loss" of an agent who reports q when the true distribution is p, as compared to to an agent who reports the true distribution p. Is is easy to verify that L_S is nonnegative everywhere if and only if S is proper.

c. (4 points). Show that a scoring rule S is strictly proper if and only if its loss function is a Bregman divergence on Δ_n.

d. (7 points). Recall from class that the loss function of the logarithmic scoring rule is the KL-divergence. One interpretation, which we mentioned in class, is that the logarithmic scoring rule incentivizes an agent to learn the true distribution as well as possible in KL-distance. It is natural to ask whether we can do the same for one of the most commonly used distance measures on distributions: the total variation distance. It turns out that this is impossible: there is no strictly proper scoring rule whose loss function is strictly increasing in the total variation distance. Since this general impossibility result is too cumbersome to prove for a homework problem, we ask you to prove a special case which captures the same ideas here: that there does not exist a strictly proper scoring rule whose loss function is proportional to the squared total variation distance. Formally, show that there does not exist a strictly proper rule $S : \Delta_n \times [n] \to \mathbb{R}$ with $L_S(p, q) = \alpha ||p - q||_1^2$ for a constant $\alpha > 0$.

Problem 2. (10 points)

Recall the peer prediction mechanism presented in class: Each agent $i \in [n]$ is designated a peer $r(i) \in [n] \setminus \{i\}$, and i's type report is treated as a (probabilistic) prediction of $r(i)$'s type report, to be rewarded using a strictly proper scoring rule. Under mild technical assumptions on the joint distribution of types and the state of nature, we showed that truthful reporting is a strict Bayes-Nash equilibrium in the peer prediction mechanism. The aforementioned technical assumptions are the following: the types are conditionally i.i.d. given the state of nature, and two different types of one agent induce two different posterior beliefs on the type of any other agent.

Despite its elegant theoretical properties, the peer prediction mechanism we saw in class suffers from a flaw which makes it impractical in some settings: truthful reporting may not be the only Bayes-Nash equilibrium. In particular, agents may coordinate on an uninformative equilibrium, perhaps as a way to avoid revealing their private types without relinquishing the monetary rewards of the mechanism.

a. (4 points). Construct an example in which the peer prediction mechanism presented in class admits an *uninformative Bayes-Nash equilibrium*: this is a Bayes-Nash equilibrium in which each agent's report does not depend on his type, despite the agent having multiple types in the support of his type distribution. Your example should obey the technical assumptions we made in class (outlined above). You will only need two agents to construct your example.

b. (6 points). When there are many agents, we can hope to carefully select the mapping $r(.)$ of agents to their peers so that such uninformative equilibria are highly unstable: if only a small fraction of the agents report truthfully, then this will incentivize the rest to report truthfully as well. In fact, we will only need a single agent to behave himself and report truthfully!

Formally, suppose there are $n \geq 2$ agents, and construct $r(.)$ so that the truthful equilibrium is the only one in which at least one agent reports truthfully. Note that $r(r(i))$ need not be equal to i. You may make the same technical assumptions we used in class (outlined above).

Problem 3. (20 points)
In this problem, we will explore a proof of the DKW inequality. We use the following notation: Let X be a random variable supported on $\{1, \dots, n\}$ and $X_1, \dots, X_m$ be independent samples of X. For $\ell = 1, \dots, n$ and $i = 1, \dots, m$, we define the random variable $X_i^{(\ell)} = \mathbf{1}_{X_i \le \ell}$. The *empirical CDF* of X with respect to the samples $X_1, \dots, X_m$ is defined as $\widehat{F}_X(\ell) = \sum_{i=1}^m X_i^{(\ell)}/m$. Let $F_X(\ell)$ denote the CDF of X, i.e., $F_X(\ell) = \mathbf{Pr}[X \le \ell]$. The DKW inequality is the following fact:

Theorem 1 (DKW Inequality)**.** *For $m = \Omega(1/\epsilon^2)$, with probability at least $9/10$ we have that*

$$\max_{1 \le \ell \le n} |\widehat{F}_X(\ell) - F_X(\ell)| \le \epsilon \, .$$

(a) **(4 points)** Use a combination of a concentration bound and the union bound to prove a weaker version of Theorem 1 for $m = \Omega(\log n/\epsilon^2)$. By adapting your previous argument, improve your bound to $m = \Omega(\log(1/\epsilon)/\epsilon^2)$.

(b) **(12 points)** Let s be the smallest integer such that $F_X(s) > 1/2$. In this part, you are asked to show that with probability at least $19/20$, it holds

$$\max_{1 \le \ell \le s-1} |\widehat{F}_X(\ell) - F_X(\ell)| \le \epsilon \, . \tag{1}$$

This proof requires a sequence of steps:

(i) Consider the differences $D_\ell = \widehat{F}_X(\ell) - F_X(\ell)$, for $1 \le \ell \le s-1$, and define the sequence of random variables M_ℓ as follows: $M_0 = 0$ and for $1 \le \ell \le s-1$

$$M_\ell = \begin{cases} \frac{D_\ell}{1 - F_X(\ell)}, & \text{if } \max_{1 \le j \le \ell-1} D_j \le \epsilon \\ D_{\ell-1}, & \text{otherwise} \, . \end{cases}$$

Similarly, consider the sequence $M_0^* = 0$ and $M_\ell^* = \frac{D_\ell}{1 - F_X(\ell)}$, $1 \le \ell \le s-1$.
Prove that both $\{M_\ell^*\}_{\ell=0,\dots,s-1}$ and $\{M_\ell\}_{\ell=0,\dots,s-1}$ are martingale sequences.

(ii) Use the martingale property to prove that $\sum_{\ell=1}^{s-1} \mathbf{E}[(M_\ell - M_{\ell-1})^2] = \mathbf{E}[M_{s-1}^2]$, and similarly that $\sum_{\ell=1}^{s-1} \mathbf{E}[(M_\ell^* - M_{\ell-1}^*)^2] = \mathbf{E}[(M_{s-1}^*)^2]$. In particular, deduce that $\mathbf{E}[M_{s-1}^2] \le \mathbf{E}[(M_{s-1}^*)^2]$.

(iii) Show that

$$\mathbf{Pr}\left[\max_{1 \le \ell \le s-1} D_\ell > \epsilon\right] \le \mathbf{Pr}\left[M_{s-1} > \epsilon\right] \le \frac{4}{\epsilon^2} \mathbf{Var}[D_{s-1}].$$

(iv) Show that $\mathbf{Var}[D_{s-1}] \le O(1/m)$.

(v) Use the above to complete the proof of (1) and establish Theorem 1.

(c) **(4 points)** Prove the following high probability version of the DKW inequality:

Theorem 2 (strong DKW Inequality)**.** *For $m = \Omega(\log(1/\delta)/\epsilon^2)$, with probability at least $1 - \delta$ we have that*

$$\max_{1 \le \ell \le n} |\widehat{F}_X(\ell) - F_X(\ell)| \le \epsilon \, .$$

(Hint: Use the appropriate concentration inequality.)